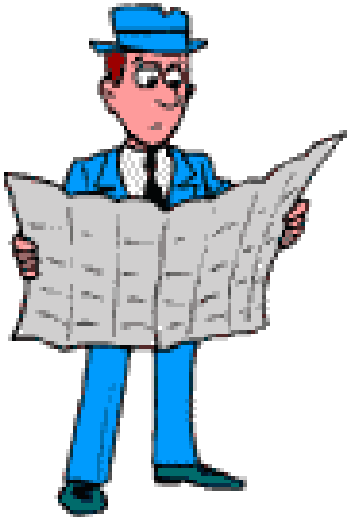




Bilgi Çıkarımı (Information Extraction-IE)



Prof.Dr. Banu Diri

Akış

- Bilgi çıkarımı nedir ?
- Mesaj anlama konferansları
- Uygulama alanları
- Yapılandırılmış, yarı yapılandırılmış dokümanlar
- NLP'nin bilgi çıkarımına katkısı
- Varlık İsmi Tanıma (NER-Name Entity Recognition)
- Kaynak seçimi
- Dinamik web sayfalarından bilgi çıkarımı
 - Alışveriş robotları (froogle)
- IE performansının ölçümü
- Bilgi çıkarımında makine öğrenmesi
 - Şablonlar metodu için bir deneme



Bilgi Çıkarım Sistemleri (IE)

- Dokümanın sınırlı ilgili bir bölümünü bulmak ve anlamak
 - Dokümanın birçok parçasından bilgi çıkarmak
 - İlgili bilginin yapısal gösterilimini sağlamak
- Bilgi tabanı (Knowledge base)

Amaç

- Bilgileri, insanların kullanacağı biçimde düzenler
- Algoritmaların yardımıyla bilgiden anlamsal çıkarımlar yapar



- Yapılandırılmamış ya da yarı yapılandırılmış dokümanlardan önceden tanımlanmış şablonlara uygun bilgileri bulma
- Yapılandırılmamış ya da yarı yapılandırılmış dokümanların yapılandırılmış veri tabanlarına dönüştürülmesi



Mesaj Anlama Konferansları

Message Understanding Conference (MUC)

- Amerikan savunma bakanlığı 1990'lardan itibaren bilgi çıkarımı konusuna eğilmiştir.
- MUC her sene yapılan bilgi çıkarımı yarışmasıdır.
- Haber makalelerinden
 - Terör olayları
 - Şirketler dünyasındaki birleşmeler, yönetim değişiklikleri konularında bilgi çıkarımı



Uygulama Alanları

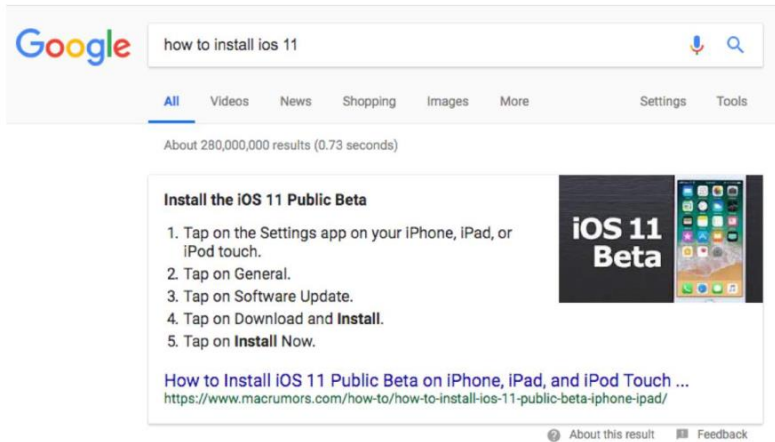
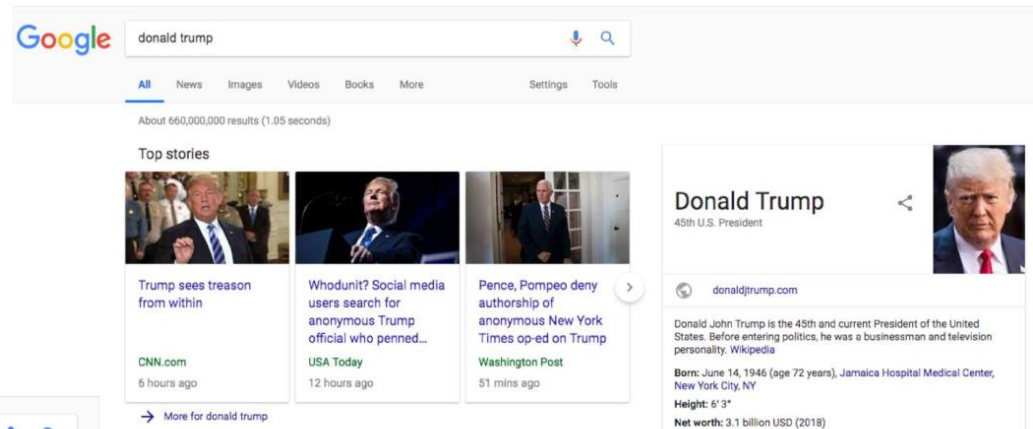
- İş ve işçi bulma
- Ürün bulma
- Seminer duyuruları
- Şirket bilgileri
- Üniversite başvuru bilgileri
- Kiralık / satılık daire, araba bilgileri
- *Ortak özellik ?*

birden fazla bilgi kaynağının araştırılması gereken durumlar



IE Teknolojisini kullanan örnekler

- Entity panel (from Google Knowledge Graph)



- How-to instructions



Doğruluk Nasıl Ölçülür ?

Hata Matrisi

	correct	not correct
selected	tp	fp
not selected	fn	tn

CoNLL: Computational Natural Language Learning
CoNLL defines the precision and the recall for named entities. This metric is considered as "**exact-match evaluation**"

Precision/Recall/F1

Precision: % of selected items that are correct $tp/(tp+fp)$

Recall: % of correct items that are selected $tp/(tp+fn)$

F-Score : Ağırlıklı harmonik ortalama

$$F = \frac{1}{\alpha \frac{1}{P} + (1-\alpha) \frac{1}{R}} = \frac{(\beta^2 + 1)PR}{\beta^2 P + R}$$

Dengelenmiş F1-Score

$\beta = 1$ (that is, $\alpha = 1/2$): $F1 = 2PR/(P+R)$



MUC : Message Understanding Conference

In MUC, detailed evaluation metrics are presented. This metric is considered as "**relaxed-match evaluation**"

Correct (COR), Incorrect (INC), Partial (PAR), Missing (MIS),
Spurious (SPU)

COR ve INC eşleşmeler kesin doğru ve yanlış

PAR tahmin edilen ile gerçek aynı değil, sınırda benzerlikler mevcut

MIS doğru olan tespit edilememiş

SPU gerçek olmayan bir şey tespit edilmiş

$$Precision = (COR + 0.5 * PAR) / (COR + SPU + 0.5 * PAR)$$

$$Recall = (COR + 0.5 * PAR) / (COR + MIS + 0.5 * PAR)$$



Yarı Yapılandırılmış Doküman Örnek İş İlanı

Subject: **US-TN**-SOFTWARE PROGRAMMER

Date: **17 Nov 1996** 17:37:29 GMT

Organization: Reference.Com Posting Service

Message-ID: <**56nigp\$mrs@bilbo.reference.com**>

SOFTWARE PROGRAMMER

Position available for Software Programmer experienced in generating software for PC-Based **Voice Mail** systems. Experienced in **C** Programming. Must be familiar with communicating with and controlling voice cards; preferable Dialogic, however, experience with others such as Rhetorix and Natural Microsystems is okay. Prefer **5** years or more experience with **PC** Based **Voice Mail**, but will consider as little as **2** years. Need to find a Senior level person who can come on board and pick up code with very little training. Present Operating System is **DOS**. May go to **OS-2** or **UNIX** in future.

Please reply to:

Kim Anderson

AdNET

(901) 458-2888 fax

kimander@memphisonline.com



Elde edilen iş özeti

computer_science_job

id: 56nigp\$mrs@bilbo.reference.com

title: SOFTWARE PROGRAMMER

salary:

company:

recruiter:

state: TN

city:

country: US

language: C

platform: PC \ DOS \ OS-2 \ UNIX

application:

area: Voice Mail

req_years_experience: 2

desired_years_experience: 5

req_degree:

desired_degree:

post_date: 17 Nov 1996



Yapılandırılmamış Doküman

Örnek Haber Metni

- 21 yaşındaki inşaat işçisi Kemal Yaprak, evine dönerken para meselesi yüzünden tartıştığı arkadaşı Hilmi Baker tarafından bıçaklanarak öldürüldü.
- Katil: Hilmi Baker
- Kurban: Kemal Yaprak
- Sebep: Para meselesi
- Suç aleti: Bıçak



Yapılandırılmış Doküman - Amazon Kitap Sayfası

....

</td></tr>

</table>

<b class="sans">The Age of Spiritual Machines : When Computers Exceed Human Intelligence

by <a href="/exec/obidos/search-handle-url/index=books&field-author=

Kurzweil%2C%20Ray/002-6235079-4593641">

Ray Kurzweil

List Price: \$14.95

Our Price: \$11.96

You Save: \$2.99

(20%)

<p>
...



Elde edilen kitap bilgileri

Title: The Age of Spiritual Machines :
When Computers Exceed Human Intelligence

Author: Ray Kurzweil

List-Price: \$14.95

Price: \$11.96

:

:



Öğelerine ayrılmış metinler

- “ye” fiilinin nesneleri yiyecek olarak sınıflandırılabilir.

geyiq.com/forum - Taksim borsada bira ile yarım döner **yerken**

geyiq.com/forum > geyiq alanı > kıl oluyorum, dumur oldum > Taksim borsada bira ile yarım döner **yerken**. Orjinalini görmek için ...

www.geyiq.com/forum/archive/index.php/t-10219.html - 3k - [Önbellek](#) - [Benzer sayfalar](#)

[Yerken Family Grave Search](#)

... It's like we've always known each other!". - Pam from CA. Advertisement. Click Here. Search Page for Surname: **Yerken**. Name: First, Middle, **Yerken** Last. ...

www.findagrave.com/surnames/y/yerken.html - 13k - [Önbellek](#) - [Benzer sayfalar](#)

[Hürriyetim](#)

... Kelebek, 25.05.2004. Ödülü, Bush kraker **yerken** söylemeyin, ... Umarım kimse ona bu ödülü kazandığını, o kraker **yerken** söylemez' diye yanıtladı. ...

www.hurriyetim.com.tr/haber/0,,sid~436@nvid~416896,00.asp - 42k - [Önbellek](#) - [Benzer sayfalar](#)

[MİLLİYET İNTERNET - BUSINESS](#)

... Zeytin **yerken** alzheimer oluyoruz haberimiz yok. Zeytini, zehirli tekstil boyası ile veya demir sülfat gübresi ile karartıp satıyorlar. ...

www.milliyet.com.tr/2003/12/12/business/bus07.html - 30k - [Önbellek](#) - [Benzer sayfalar](#)

[TürkiyeOnline.com - Haber](#)

... sağlık. Mantar **yerken** dikkat Havalarda ısınmasıyla birlikte doğada ortaya çıkan mantarların bilinçsiz olarak tüketilmesinin, zehirlenmelere neden ...

www.turkiyeonline.com/haber/saglik/haber.php?story=2004_04_02_mantar - 20k - [Önbellek](#) - [Benzer sayfalar](#)



NLP'nin Bilgi Çıkarımına Katkısı

- Bilgiler dinamik web sayfalarından çıkarılacaksa basit regex şablonları yeterli olabilir.
- Bilgiler, insanlar tarafından yazılmış metinlerden çıkarılacaksa NLP metotları yardımcı olabilir.
 - Part-of-speech (POS) tagging
 - Kelimelerin türünü (isim, fiil, sıfat vb.) belirleme
 - Sentaktik çözümleme
 - Kelime gruplarını, ağaçları belirleme, öğeleri bulma: NP, VP, PP
 - Anlamsal Kelime Sınıfları (WordNet'den)
 - KILL: kill, murder, assassinate, strangle, suffocate
 - Name Entity Recognition
- Örnek *Öldürülen* şablonu:

Bart killed Rose.

 - Öncül şablon: [POS: V, **synset: KILL**]
 - Şablon: [**Phrase: NP**]



Varlık İsmi Tanıma (*Name Entity Recognition*) Nedir?

Bir doküman içerisindeki varlıkları tespit etmek ve onları tiplerine göre sınıflara ayırmak

[*Barack Obama*] arrived this afternoon in [*Washington, D.C.*].

[*President Obama*]'s wife [*Michelle*] accompanied him

[*TNF alpha*] is produced chiefly by activated [*macrophages*]

[*Barack Obama*] arrived this afternoon in [*Washington, D.C.*].

[*President Obama*]'s wife [*Michelle*] accompanied him

PERSON
LOCATION

[*TNF alpha*] is produced chiefly by activated [*macrophages*]

PROTEIN
CELL



• Şablon Örnek(ler)

Examples of Keywords:

person titles (e.g., Mr., Jr., Ph.D.)

company designators (e.g., Corp., Inc., Co.)

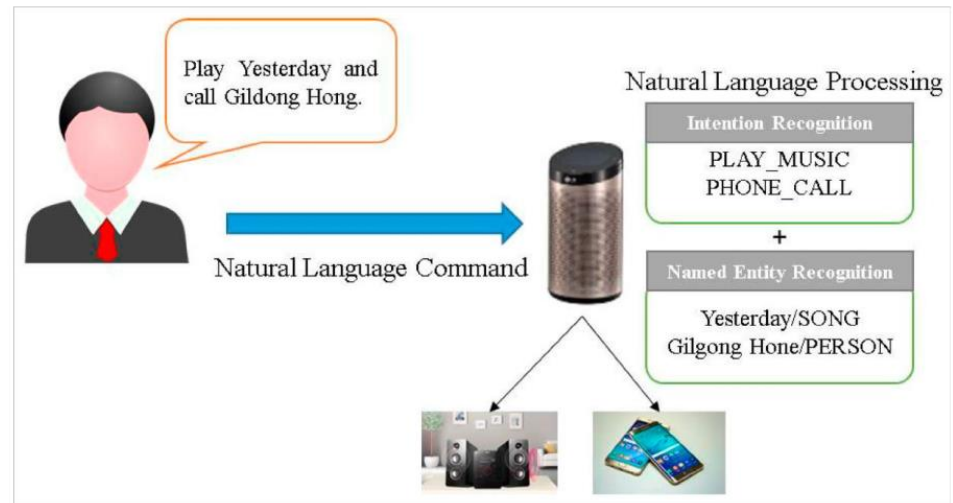
- {TITLE} {PERSON}
Ex: *"U.S. President George Bush", "Mr. Frank Leonard"*
- {PERSON}, the {TITLE} of {ORGANIZATION}
Ex: *"Fred Martin, the CEO of XYZ Corp."*
- {PERSON} joined {COMPANY}
Ex: *"Mary Smith joined Microsoft."*
- headquarters in {LOCATION}
Ex: *"headquarters in London"*
- {LOCATION}, {LOCATION}
Ex: *"Salt Lake City, Utah"*



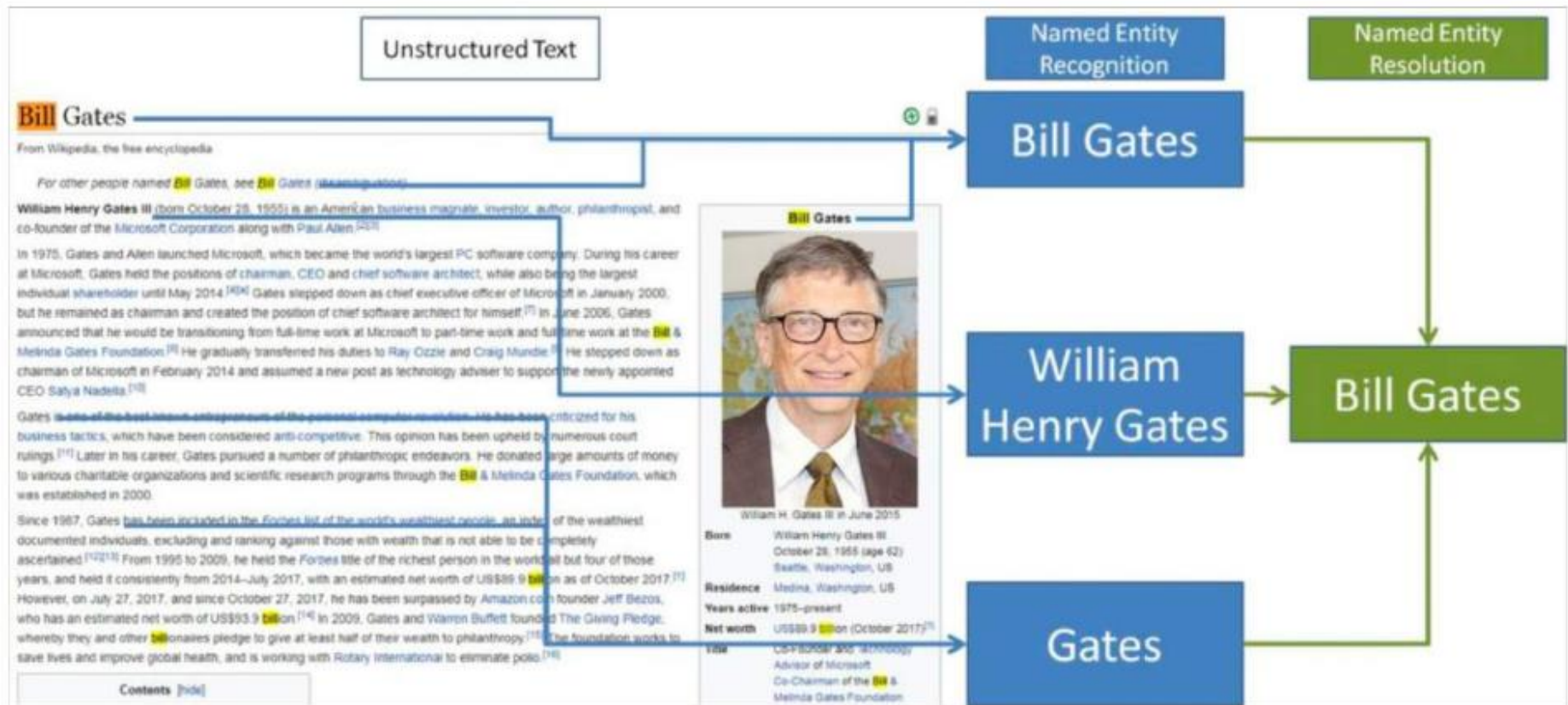
Neden NER?

- Coreference Resolution / Entity Linking
- Relation Extraction
- Knowledge Base Construction
- Web Query Understanding
- Question Answering
- ...

Doğal Dili Anlamak



Coreference Resolution / Entity Linking



NER tanımını 1995 yılında MUC-6 (Message Understanding Conference) konferansında yapılmıştır.

ENAMEX, TIMEX ve NUMEX olmak üzere 3 temel kategori ile tanımlanır

Enamex: Kişi, yer, organizasyon gibi ifadeleri

Numex: Parasal ve yüzdesel ifadeleri

Timex: Gün ve tarih gibi zamansal ifadeleri tanımlamak için kullanılmaktadır.



Public Benchmarks for NER

- **CoNLL-2002 and CoNLL-2003 (British newswire)**
 - Multiple languages: Spanish, Dutch, English, German
 - 4 entities: Person, Location, Organization, Misc
- **MUC-6 and MUC-7 (American newswire)**
 - 7 entities: Person, Location, Organization, Time, Date, Percent, Money
- **ACE**
 - 5 entities: Location, Organization, Person, FAC, GPE
- **BBN (Penn Treebank)**
 - 22 entities: Animal, Cardinal, Date, Disease, ...



NER için kullanılan 3 yaklaşım

- Kural Tabanlı Yaklaşımlar (Rule Based)
- Standart Sınıflandırıcılar (KNN, Decision Tree, Naïve Bayes, SVM, ...)
- İstatistiksel Yöntemler (HMM, CRF)
- Hibrit Yöntemler
- Sinir Ağı (Neural Network) Tabanlı Yaklaşımlar
- Transformer Tabanlı Yaklaşımlar



■ **Kural Tabanlı Yaklaşımlar (Rule Based)**

- Sisteme insan tarafından tanımlanan kuralların verilmesi gerekir
- Maliyetlidir, tasarımı kolay değildir
- Bir alan için çıkarılan kurallar başka bir alan için uyarlanamaz

Düzenli ifadelerden çıkarım:

- Telephone number
- E-mail
- Capitalized names

Knowledge Engineering



- + very precise (hand-coded rules)
- + small amount of training data
- expensive development & test cycle
- domain dependent
- changes over time are hard

- matches valid phone numbers like 900-865-1125 and 725-1234
- incorrectly extracts social security numbers 123-45-6789
- fails to identify numbers like 800.865.1125 and (800)865-CARE

RegEx = $([d\{3\}[-.\ ()])\{1,2\}[\dA-Z]\{4\}$

Location

- Capitalized word + {city, center, river} indicates location

Ex. *New York city*

Hudson river

- Capitalized word + {street, boulevard, avenue} indicates location

Ex. *Fifth avenue*



Yapısal olmayan bir doküman için NLP den yardım alınabilir

- Part-of-speech (POS) tagging
Mark each word as a noun, verb, preposition, etc.
- Syntactic parsing
Identify phrases: NP, VP, PP
- Semantic word categories (e.g. from WordNet)
KILL: kill, murder, assassinate, strangle, suffocate



Kural Tabanlı Yöntemler Her Zaman Çalışmaz

- Özel isimler büyük harf ile başlar ama cümle de büyük harfle başlar
- Web sayfalarındaki başlıkların hepsi büyük harf yazılabilir
- Elimizde *movie titles, books, singers, restaurants, etc.* olmayabilir
Gazetteer ihtiyaç duyulur
- Özel isimlerde karışıklık olabilir
Jordan the *person* vs. Jordan the *location*
JFK the *person* vs. JFK the *airport*
May the *person* vs. May the *month*
- Bir ismin farklı yazılımları olabilir
Xiang Ren
Prof. Ren
Dr. Ren
Xiang



Standart Sınıflandırıcılar

Eğitim-Training

1. Eğitim için kullanılacak temsili dokümanlar toplanır
2. Her bir varlık ismi etiketlenir
3. Metin ve sınıflara uygun özellikler çıkarılır
4. İşaretlenmiş cümlelerden her varlık isminin sınıfını tahmin edecek sınıflandırıcı ile veri eğitilir

Test-Testing

1. Test edilecek doküman seçilir
2. Eğitilmiş model çalıştırılır
3. Her varlık için uygun etiket belirlenir



NER için klasik öğrenme yöntemleri

- KNN
- Decision Tree
- Naive Bayes
- SVM
- ...
- Boosting



Etiketleme Formatları

IOB, IOB1, IOB2, BILOU (B, I, O, L ve U)

B → Bir varlık isminin başladığını gösterir (Begin)

I → Varlık isminin devam ettiğini gösterir (Inside)

L → Varlık isminin son kelimesi olduğunu gösterir (Last)

O → Herhangi bir kategoriye ait olmayan varlık isimleri için kullanılır (Other)

U → Tek kelimelik varlık isimlerini tanımlamak için kullanılır

IOB formatı

Mustafa	B-Person
Kemal	I-Person
Atatürk	I-Person
kurtuluş	O
mücadeles	O
başlatmak	O
üzere	O
Bandırma	O
vapuru	O
ile	O
19	B-Date
Mayıs	I-Date
1919	I-Date
yılında	I-Date
Samsun	B-Loc
,	O
a	O
gitmiştir.	O
.	O



K-NN

	isPersonName	isCapitalized	isLiving	teachesCS544
Jerry Hobbs	1	1	1	1
USC	0	1	0	0
eduard hovy	1	0	1	1
Kevin Knight	1	1	1	1

$$d(\text{JerryHobbs}, \text{USC}) = \sqrt{(1^2 + 0 + 1^2 + 1^2)} = 1.73$$

$$d(\text{JerryHobbs}, \text{eduard hovy}) = \sqrt{(0 + 1^2 + 0 + 0)} = 1$$

$$d(\text{JerryHobbs}, \text{Kevin Knight}) = \sqrt{(0 + 0 + 0 + 0)} = 0$$



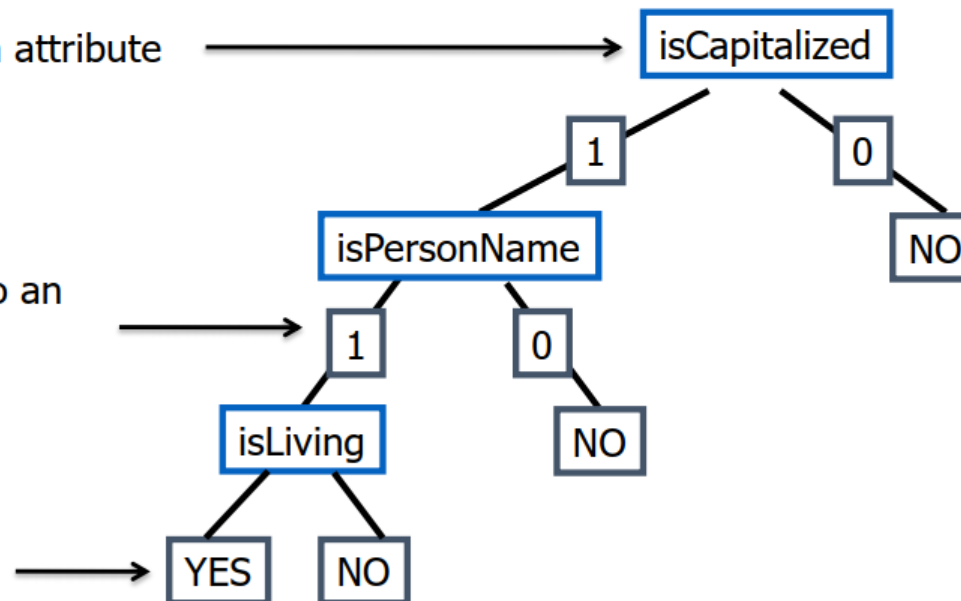
Karar Ağaçları

	isPersonName	isCapitalized	isLiving	X is PersonName?
profession	0	0	0	NO
Jerry Hobbs	1	1	1	YES
USC	0	1	0	NO
Jordan	1	1	0	NO

Each internal node tests an attribute

Each branch corresponds to an attribute value node

Each leaf node assigns a classification



NER için İstatistiksel Öğrenme Yöntemleri

- HMMs (Hidden Markov Models)
- CRFs (Conditional Random Fields)
- ...

NER için Hibrit Öğrenme Yöntemleri

- Kural Tabanlı yöntemler ile İstatistiksel yöntemler birleştirilerek kullanılır

NER için Sinir Ağı Tabanlı Öğrenme Yöntemleri

- LSTM, BiLSTM, ...



Data	Study	Method	F1 (%)
News Articles [1]	[2]	BERTurk-CRF	95.95
	[19]	Deep-BiLSTM	93.69
	[14]	BiLSTM	93.37
	[11]	CRF	91.94
	[1]	HMM	91.56
	[10]	CRF	88.94
Twitter Dataset [3]	[4]	BiLSTM-CRF	67.39
	[25]	Reg. Avg. Perp.	48.96
	[7]	Rule-based	38.01
	[3]	CRF	19.28
	[23]	WAN	15.43

[1] Gökhan Tür, Dilek Hakkani-Tür, and Kemal Oflazer. 2003. A statistical information extraction system for Turkish. *Natural Language Engineering* 9, 2 (2003), 181–210.

[2] Aras, Gizem, et al. "An evaluation of recent neural sequence tagging models in Turkish named entity recognition." *Expert Systems with Applications* 182 (2021): 115049.

[3] Gökhan Çelikkaya, Dilara Torunoğlu, and Gülşen Eryiğit. 2013. Named entity recognition on real data: a preliminary investigation for Turkish. In *2013 7th International Conference on Application of Information and Communication Technologies*. IEEE, 1–5.

[4] Emre Kağan Akkaya and Burcu Can. 2021. Transfer learning for Turkish named entity recognition on noisy text. *Natural Language Engineering* 27,1 (2021), 35–64.

[7] Dilek Küçük and Ralf Steinberger. 2014. Experiments to Improve Named Entity Recognition on Turkish Tweets. In *Proceedings of the 5th Workshop on Language Analysis for Social Media (LASM)*. 71–78.

[10] Reyhan Yeniterzi. 2011. Exploiting Morphology in Turkish Named Entity Recognition System. In *Proceedings of the ACL 2011 Student Session*. Association for Computational Linguistics, Portland, OR, USA, 105–110

[11] Gökhan Akın Şeker and Gülşen Eryiğit. 2012. Initial explorations on using CRFs for Turkish named entity recognition. In *Proceedings of COLING2012*. 2459–2474.

[14] Onur Güngör, Suzan Üsküdarlı, and Tunga Güngör. 2018. Recurrent neural networks for Turkish named entity recognition. In *2018 26th Signal Processing and Communications Applications Conference (SIU)*. IEEE, 1–4.

[19] Asim Güneş and A Cüneyd Tantuğ. 2018. Turkish named entity recognition with deep learning. In *2018 26th Signal Processing and Communications Applications Conference (SIU)*. IEEE, 1–4.

[23] Onal, Kezban Dilek, and Pinar Karagoz. "Named entity recognition from scratch on social media." *Proceedings of 6th International Workshop on Mining Ubiquitous and Social Environments (MUSE)*, co-located with the ECML PKDD. Vol. 104. 2015. Stefan Schweter. 2020. BERTurk - BERT models for Turkish.

[25] Eda Okur, Hakan Demir, and Arzucan Özgür. 2016. Named Entity Recognition on Twitter for Turkish using Semi-supervised Learning with Word Embeddings. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*. 549–555.



NER ile Bilgi Çıkarımı

Giriş: Web sayfaları

NER

Şablon Eşleme

Çıkış: İkililer

Brent Barlow, a software analyst and beta-tester at Apple Computer's headquarters in Cupertino, was fired Monday for "thinking a little too different."

doc4

<PERSON>Brent Barlow</PERSON>,
a software analyst and beta-tester at
<ORGANIZATION>Apple Computer</ORGANIZATION>'s
headquarters in <LOCATION>Cupertino</LOCATION>, was fired
Monday for "thinking a little too different."

doc4

<ORGANIZATION> = Apple
Computer
<LOCATION> = Cupertino
Pattern = p1

doc4

Extraction Patterns

<ORGANIZATION>'s p1
headquarters in <LOCATION>

<ORGANIZATION>,
based in <LOCATION> p2

tid	Organization	Location	W
1	Eastman Kodak	Rochester	0.9
2	Apple Computer	Cupertino	0.8

Useful
doc2
doc4

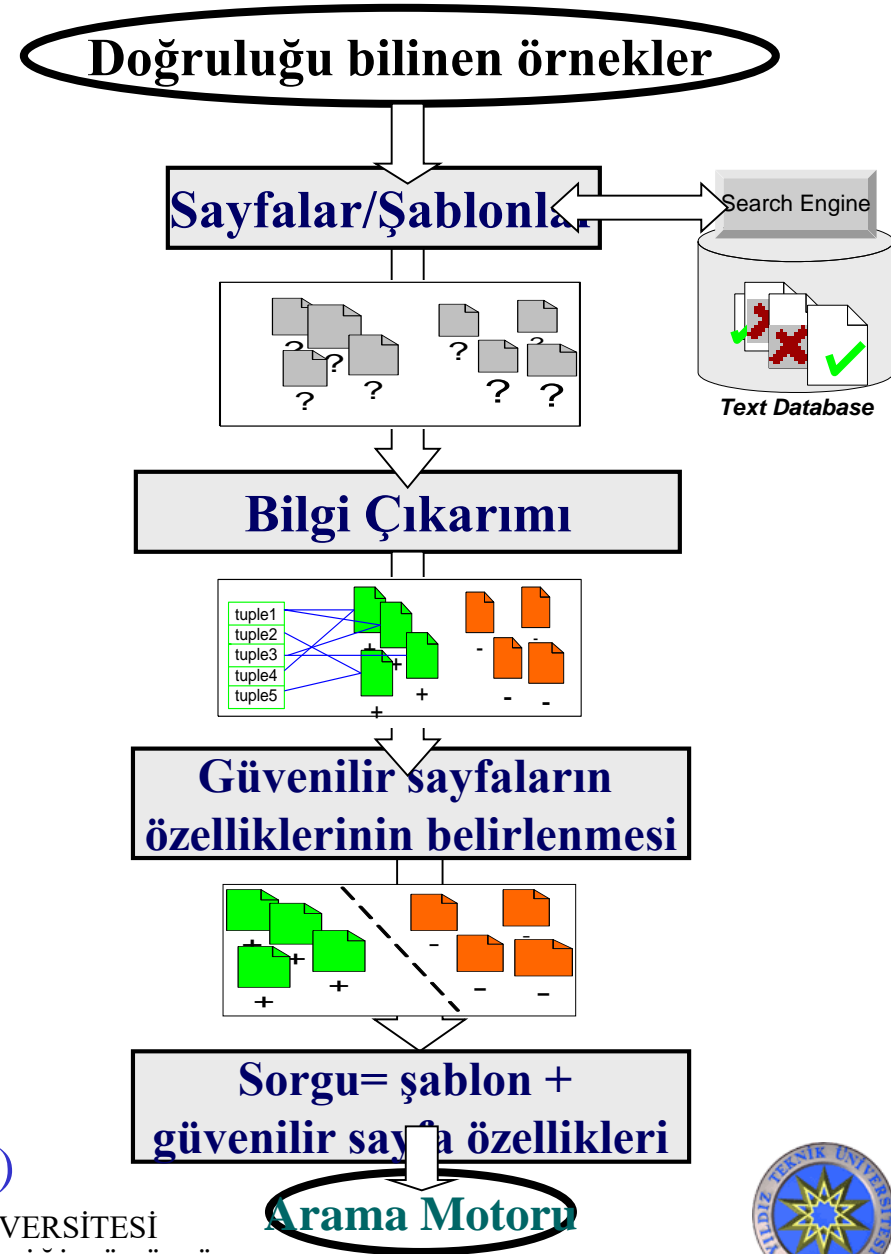


Kaynak Seçimi

- Doğruluğu bilinen örnekler arama motoruna gönderilir.
- Sonuçlardan şablonlar çıkarılır.
- Bu şablonlar arama motoruna gönderilir, uyan sayfalardan bilgiler çıkarılır.
- Çıkarılan bilgilerin doğruluğu bir veritabanından kontrol edilir.
- Doğru ve yanlış bilgi çıkarılan web sayfaları işaretlenir.
- Bu sayfaların özellikleri çıkarılır.
- Bundan sonra şablonlarla birlikte doğru sayfaların özellikleri de aratılır.
- Bu sayede sadece güvenilir sayfalarda arama yapılmış olunur.


ÖZETLE: Bulunan şablonlara ek olarak, güvenilir sayfaların özellikleri de bulunarak sorguya eklenir.

SAYFA ÖZELLİKLERİ: İçinde geçen kelimeler, url'sinde geçen kelimeler (ör: *edu*)




Dinamik Web Sayfalarından Bilgi Çıkarım Metotları


- Birçok web sayfası veritabanlarından dinamik olarak oluşturulur.
- Dinamik web sayfalarında html tag'leri tekrar eder.
- Tekrar eden kalıplar arasında aynı tür bilgiler yer alır.



Hot prices and selection in our Electronics store!




Software running slow? Get more memory!



Visit our Computer & Internet Books Section





So You'd Like to... Offer your advice





LOOK INSIDE

- 


1. **WebObjects 5 for Java: A Developer's Guide (With CD-ROM)**
by Jesse Feiler (**Paperback**)
Usually ships in 24 hours
List Price: \$59.99
Our Price: **\$59.99**
Add to cart
- 

2. **Mastering Windows XP Professional**
by Mark Minasi (**Paperback** - October 2001)
Avg. Customer Rating: ★★★★★
Usually ships in 24 hours
List Price: \$39.99
Our Price: **\$27.99**
You Save: **\$12.00** (30%)
Add to cart
Or buy used: \$25.19
- 

3. **LOOK INSIDE CCNA Virtual Lab, Gold Edition**
by Todd Lammle, et al (**CD-ROM**)
Avg. Customer Rating: ★★★★★
Usually ships in 24 hours
List Price: \$149.99
Our Price: **\$104.99**
You Save: **\$45.00** (30%)
Add to cart
Or buy used: \$80.00
- 

4. **Microsoft Windows 2000 MCSE Core Requirements Training Kit (With CD-ROM)**
by Microsoft Corporation (Editor) (**Paperback**)
Avg. Customer Rating: ★★★★★
Usually ships in 24 hours
List Price: \$199.99
Our Price: **\$139.99**
You Save: **\$60.00** (30%)
Add to cart
Or buy used: \$19.99
- 


5. **Microsoft Windows Xp Inside Out**
by Ed Bott, et al (**Paperback**)
Avg. Customer Rating: ★★★★★
Usually ships in 24 hours
List Price: \$44.99
Our Price: **\$31.99**
Add to cart




Have To Have Cisco Books! A list by newccnp2001, Recent CCNA/CCDA/CCNP certs (9 item list)




Windows 2000 Pro: A list by Teri Kieffer, editor, Amazon.com (13 item list)



Great Computer Books: A list by jeffloft, Computer Enthusiast (25 item list)



Pragmatic Programmer Suggested Reading: A list by Andrew D Robinson, Internet Project Manager (15 item list)



Great Programming Books: A list by brookeg, Editor, Amazon.com (15 item list)

Tablomuzun Satırlarını Belirlemek

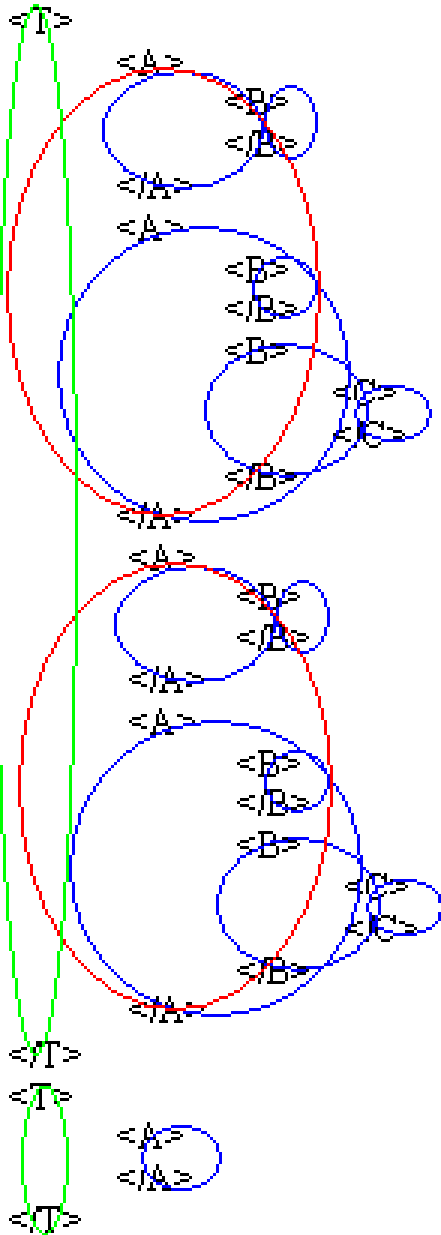
Satırlar başlayıp biten HTML tag'lerinden oluşur.

Hangi tag'le satırın başlayıp bittiğini bulmak önemli.

Kural 1: Her satırdaki HTML tag sayısı birbirine yakındır/eşittir.

Kural 2: En fazla tag içeren tekrarlı çevrim satırı gösterir.





- Yanda olası tüm satırlar gözükmemektedir.
- Her satırda yakın sayıda tag olması şartından dolayı T'lerin satırları oluşturmadığı görülür.
- En fazla sayıda tag içeren satır seçileceğinden kırmızı ile gösterilen kısımlar satırlar olarak belirlenecektir.

Alışveriş Robotları

- Tekrarlı HTML tag'leri kullanılarak bilgi çıkarılan sistemlere örnek olarak çeşitli web sitelerinde satılan ürünlerin bilgilerini tek bir sayfada toplayan sistemler verilebilir.
- Örnek Siteler:
 - MySimon
 - Cnet
 - BookFinder
 - Froogle



Alışveriş/Haber Toplama Robotlarının Çalışma Adımları

- 1- Her satıcı/haberci site bilgi çıkarım mekanizmasını kurar.
- 2- Kullanıcıdan sorgusunu alır (tür, fiyat vs.).
- 3- Her site için:
 - Kullanıcı sorgusu siteye gönderilir.
 - Sonuç sayfaları alınır.
 - Sonuç sayfası, o sayfanın bilgi çıkarım mekanizmasıyla işlenir. Sonuçlar kendi veritabanına kaydedilir.
- 4- Sonuçlar (fiyatlara/tarihlere göre) sıralanır.
- 5- Sonuçlar HTML formatına çevrilir. Kullanıcıya döndürülür.



Şablonların bulunması

- Keşfetmek istediğimiz ikililerin aralarındaki ilişki türü belirlenir. Ör: “Tüm X’ler Y’dir”.
- Bilinen X,Y ikilileri Google’da aratılır.
- X ve Y arasındaki şablonlar ve frekansları belirlenir.
- En yüksek frekansa sahip şablonlar bu ilişki türünün şablonları olurlar.



Bulunan şablonlardan örnekler tüm X'ler Y'dir için

- ve diğer
- ler ve diğer
- ve benzeri
- veya diğer
- türü olan
- ları ve diğer
- lar ve diğer
- ve her türlü
- lerden biri olan
- leri ve diğer
- larından biri olan
- lerinden biri olan
- lardan biri olan
- adı olan
- ve her tür



Bulunan şablonlardan örnekler

X'in yeri Y'dir için

- y deki x
- y de bulunan x
- y de x
- x y de
- x y ili sınırları içerisindedir
- y ili sınırlarında kalan x
- y ili sınırları içinde bulunan x
- y ilçesi sınırları içinde bulunan x
- x y nin sınırları içerisindedir
- x/y
- x / y
- $x-y$
- x y ye zz km

x, y ye zz km

x (y ye zz km

x, y

$x - y$

x bulunduğu yer:y

$y-x$

$x(y$

$x(y)$



Şablonlara uygun ikililerin bulunması

- Google'da bulunan şablonlar aratılır.
- Sonuç sayfalarındaki şablonların sağ ve sollarındaki kelimeler alınır ve bir dosyaya kaydedilir.



Şablonlara uygun ikililerden örnekler

Tüm X'ler	Y'dir
• kontrolör	personel
• teçhizat	malzeme
• kemer	teçhizat
• protein	gıda
• Azerbaycan	bölge
• Ceyda	yardımcı
• komünizm	ideoloji
• delta	Gediz
• kurum	Kocaelispor
• fotoğrafçı	Robert

- tür flamingo
- ünite aksesuar
- bedel masraf
- din azınlık
- çelik yapı
- yem araç
- kız sıfat
- yapı sorun
- ölçü şart



İkililerin elle sınıflandırılması

- Bulunan ikililerden hangilerinin “Tüm X’ler Y’dir” ilişkisine sahip olup olmadığı elle işaretlenir.



Kaynaklar

- Rada Mihalcea, “NLP lecture slides”
- www.ccs.neu.edu/home/futrelle/bionlp/psb2001/Hawaii-Tutorial-Tsujii.ppt
- www.cs.utexas.edu/users/mooney/ir-course/slides/InformationExtraction.ppt
- www.cs.columbia.edu/~eugene/talks/icde2003.ppt
- www.isi.edu/natural-language/teaching/cs544/cs544-9-apr04.ppt
- www.cs.sfu.ca/~zshi1/personal/projects/Presentation_thesis.ppt

